Zero-Shot Scene Reconstruction from Single Images with Deep Prior Assembly

Junsheng Zhou[1] **Yu-Shen Liu**[1†] **Zhizhong Han**[2]
School of Software, Tsinghua University, Beijing, China[1]
Department of Computer Science, Wayne State University, Detroit, USA[2]
zhou-js24@mails.tsinghua.edu.cn liuyushen@tsinghua.edu.cn h312h@wayne.edu

Abstract

Large language and vision models have been leading a revolution in visual computing. By greatly scaling up sizes of data and model parameters, the large models learn deep priors which lead to remarkable performance in various tasks. In this work, we present deep prior assembly, a novel framework that assembles diverse deep priors from large models for scene reconstruction from single images in a zero-shot manner. We show that this challenging task can be done without extra knowledge but just simply generalizing one deep prior in one sub-task. To this end, we introduce novel methods related to poses, scales, and occlusion parsing which are keys to enable deep priors to work together in a robust way. Deep prior assembly does not require any 3D or 2D data-driven training in the task and demonstrates superior performance in generalizing priors to open-world scenes. We conduct evaluations on various datasets, and report analysis, numerical and visual comparisons with the latest methods to show our superiority. Project page: https://junshengzhou.github.io/DeepPriorAssembly.

1 Introduction

Reconstructing scenes from images is a vital task in 3D computer vision and computer graphics. It bridges the gap between the 2D images that can be easily captured by phone cameras and the 3D geometries of scenes for various real-world applications, e.g., autonomous driving, augmented/virtual reality and robotics. Reconstructing scenes from multi-view images [60, 67] is well-explored to recover 3D geometries with multi-view consistency and camera poses. However, reconstructing a scene from a single RGB image is still challenging, which is extremely difficult due to inadequate information. Recent works [10, 42] try to solve this task as a reconstruction problem which leverages neural networks with an encoder-decoder architecture to draw supervisions from pairs of images and 3D ground truth geometries and layouts. Nevertheless, due to the limited amount of image-scene pairs, these methods struggle to generalize to out-of-distribution images in open world.

Large language and vision models have been extensively studied in the past few years, which revolutionized neural language processing [58, 6], 2D/3D representation learning [15, 77] and content generation [51, 26], etc. By greatly scaling up sizes of training samples and model parameters, large models show brilliant capabilities and remarkable performance. However, they are limited in a specific task, which limits their capability in high level perception tasks. Driven by the observation, we raise an interesting question: can we assemble series of deep priors from large models, which are experts with different modalities in different tasks, to solve an extremely challenging task that none of them can accomplish alone?

[†] The corresponding author is Yu-Shen Liu.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1234

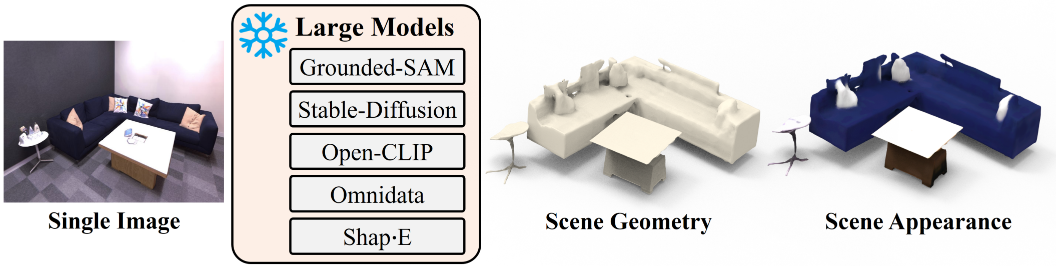

Figure 1: **An illustration of our work.** We assemble diverse deep priors from large models with frozen parameters for scene reconstruction from single images in a zero-shot manner.

In this work, we propose *deep prior assembly*, a novel framework which assembles diverse deep priors from large models for scene reconstruction from single images in a zero-shot manner. We rethink this task from a new perspective, and decompose it into a set of sub-tasks instead of seeking to a data-driven solution. We narrow down the responsibility of each deep prior on a sub-task that it is good at, and introduce novel methods related to poses, scales, and occlusion parsing to enable deep priors to work together in a robust way.

Specifically, we first detect and segment the instances in the input image with Grounded-SAM [29, 33], which is a variation of Segment-Anything Model [29]. For the segmented instances that are often corrupted due to occlusions or of low resolution, we leverage Stable-Diffusion [51] to enhance and inpaint images containing the segmented instances. However, the Stable-Diffusion often produces some predictions which drift away from the input instances and do not conform to the original appearance. To solve this issue, we introduce to use Open-CLIP [47, 23] to filter out the bad samples and select the ones matching the input instance most. We then generate the 3D models for each instance with Shap·E [26] using the amended instance images as input. Additionally, we estimate the depth of the origin image with Omnidata [13] as the 3D scene geometry prior. To recover a layout consistent to the image, we propose an approach to optimize the location, orientation and size for each 3D instance to fit it to the estimated segmentation masks and the depths.

Deep prior assembly merely generalizes deep priors and does not require additional data-driven training for extra prior knowledge. Our evaluations on various open-world scenes show our capability of reconstructing diverse objects and recovering plausible layout merely from a single view angle. Our main contributions can be summarized as follows.

- We propose deep prior assembly, a flexible framework that assembles diverse deep priors from large models together for reconstructing scenes from single images in a zero-shot manner.
- We introduce a novel approach to optimizing the location, orientation and scale of instances by matching them with both 2D and 3D supervision.
- We evaluate deep prior assembly for generating diverse open-world 3D scenes, and show our superiority over the state-of-the-art supervised methods.

2 Related Work

2.1 Large Models in Different Modalities

Recently, it has been drawing significant attention on scaling up deep models for much more powerful representations and higher performances with different modalities (e.g. NLP, 2D vision). Starting from NLP, recent works in scaling up pre-trained language models [6, 34, 48] have largely revolutionized natural language processing. Some researches translates the progress from language to 2D vision [47, 12, 3, 20, 15] and 3D vision [77] via model and data scaling.

Except for the large foundation models which focus on producing large-scale representations for language, 2D images or 3D point clouds, some researches explore large models for specific tasks (e.g. text-to-image generation [51], image segmentation [29], 3D analysis [62, 78, 76, 31, 30] and 3D object generation [80, 61, 35, 79]) and have shown remarkable performance. Stable Diffusion trains a large model of latent diffusion models and achieves commercially available 2D content generation

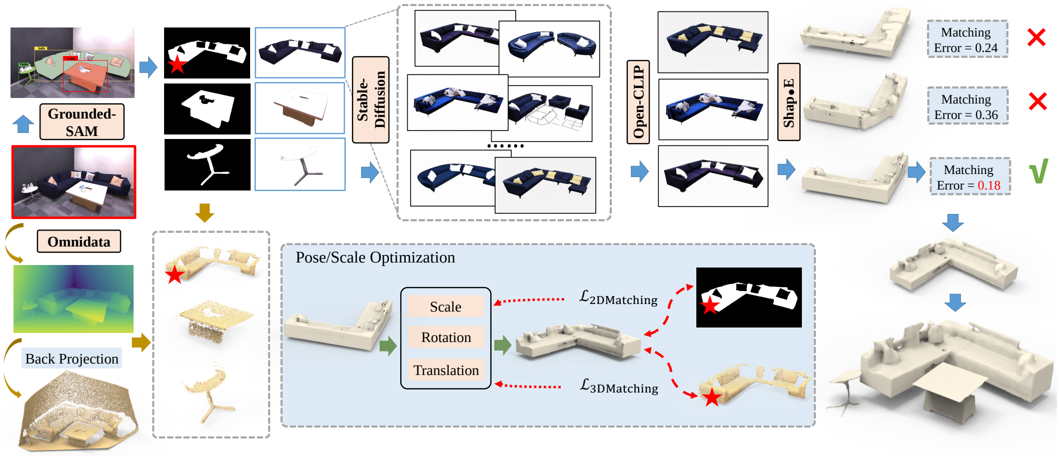

Figure 2: **The overview of deep prior assembly.** Given a single image of a 3D scene, we detect the instances and segment them with Grounded-SAM. After normalizing the size and center for the instances, we attempt to amend the quality of the instance images by enhancing and inpainting them. Here, we take a sofa in the image for example. Leveraging the Stable-Diffusion model, we generate a set of candidate images through image-to-image generation, with additional guidance from a text prompt of the instance category predicted by Grounded-SAM. We then filter out the bad generation samples with Open-CLIP by evaluating the cosine similarity between the generated instances and original one. After that, we generate multiple 3D model proposals for this instance with Shap·E from the Top-K generated instance images. Additionally, we estimate the depth of the origin input image with Omnidata as a 3D geometry prior. To estimate the layout, we propose an approach to optimize the location, orientation and scale for each 3D proposal by matching it with the estimated segmentation masks and the depths (the $\star$ for the example sofa). Finally, we choose the 3D model proposal with minimal matching error as the final prediction of this instance, and the final scene is generated by combining the generated 3D models for all detected instances.

effects. Segment Anything Model (SAM) [29] revolutionize the field of image segmentation by training models with large-scale annotated data. Omnidata [13] trains the large depth estimation model with various data sources for bringing robust 3D awareness to pure RGB images. In the 3D domain, the recent works Point·E [40] and Shap·E [26] collect millions of 3D objects to train large 3D models for generating 3D geometries from rendering-style images or texts. In this work, we aim at leveraging the powerful capabilities of the large models in different modalities and different domains to solve a challenging task, i.e. scene generation from single images, by assembling deep priors together.

2.2 Scene Reconstruction from Images

Recovering the underlying 3D surfaces of scenes from images [60, 70, 22, 19, 41] or point clouds [72, 75, 73, 25, 36, 74, 24, 43] is a long-standing task in 3D computer vision. Most of the previous works focus on the multi-view reconstruction with the input dense images captured around the scene. Classic multi-view stereo methods [1, 5, 4] mainly represent the scene by estimating depths for dense images with feature matching. Inspired by NeRF [39] which performs volume rendering for scene representation, a series of works [60, 44, 65, 59, 68, 69] introduce the neural implicit surface reconstruction by learning signed distance fields [45] or occupancy fields [38] for scenes from multi-view images. NeuRIS [59] proposes to use normal priors for indoor scene reconstruction, and MonoSDF [67] further introduces monocular depth cues for improving scene geometries.

3 Method

Overview. The overview of deep prior assembly is shown in Fig. 2. We will start from an introduction of the task decomposition in Sec. 3.1 and then present the pipeline for solving each of the decomposed sub-tasks using a deep prior from a specific large model in Sec. 3.2. Finally, we introduce an optimization-based approach for layout estimation in Sec. 3.3.

3.1 Task Decomposition

Revealing 3D scene geometries from a single image is an extremely challenging task duo to limited context and supervisions. Instead of using a data-driven strategy to learn priors [42, 10], we reformulate this task from a new perspective. We decompose it into a set of sub-tasks, each of which can be done using one deep prior without a need of learning extra knowledge. More specifically, we can progressively resolve the task by:

1) First, performing detection and segmentation on the input image to acquire the segmentation images, masks and category labels for all detected instances.
2) Amending instance images through enhancing and inpainting to improve the image qualities.
3) Generating a set of 3D model proposals for each instance from 2D segmented images.
4) Estimating the layout by predicting the location, rotation, and scale for each 3D proposal to put them to the correct position of the 3D scene.
5) Producing a scene reconstruction by applying the estimated layout and shape poses with reconstructed instances.

3.2 Assembling Large Models

Inspired by the remarkable performances of recent large models, we propose to assign an expert large model in each sub-task, which maximizes their abilities for modeling a scene in a zero-shot manner.

Detect and Segment 2D instances. To reveal a scene S from a single image I, we first detect the instances in I and separate multiple objects into single instances. In this way, we can reconstruct a scene at shape level, which simplifies the task.

Mask R-CNN vs. SAM vs. Grounded SAM. Detecting and segmenting images have been widely explored in the past few years. Mask R-CNN [21] is widely adopted as a popular and robust backbone. However, the performance of Mask R-CNN does not generalize well since it is only trained under a relative small dataset. Recently, the large SAM [29] have shown promising segmentation accuracy by scaling up parameters and using more training samples, nevertheless, it only predicts fine-grained masks but with few semantic concepts. Thus, we adopt Grounded-SAM, which is an improved version of SAM by introducing Grounding-DINO [33] as an open-set detector and using SAM to jointly predict detection boxes, segmentation masks and category labels for each instance, formulated as:

$$\{m_i, c_i, d_i\}_{i=1}^{N} = \mathrm{GroundedSAM}(I), \text{ where } d_i > \sigma \tag{1}$$

$\{m_i, c_i, d_i\}$ includes the predicted mask m_i, category c_i and the detection confidence score d_i for the i-th instance, and N is the number of instances in this scene. We only keep the predictions with a high confidence score larger than a threshold σ. The low confident instances often contain large occlusions or wrong category predictions.

Enhance and Inpaint 2D instances. With the predicted masks $\{m_i\}_{i=1}^{N}$, we achieve the segmented instance images $\{t_i\}_{i=1}^{N}$ by masking the original input image I. We then normalize each instance image by centering it at the origin and normalizing its scale in $\{t_i\}_{i=1}^{N}$ to 0.6 of the max width or height of I. As shown in Fig. 2, the segmented instance images often suffer from occlusions or low-resolution of small instances. The low-quality images have a large negative impact on the followup 3D generation. Therefore, we propose to first improve the quality of instance images by enhancing and inpainting them with the large model Stable-Diffusion [51].

Specifically, we adopt the image-to-image generation [37] from Stable-Diffusion. We take the instance image t_i as the initialization, and add noises on it and then subsequently denoise the noise corrupted image to increase the realism through the guidance of the text prompt description from the predicted categories c_i. We find the prompt template 'High quality, authentic style {category}' works fine for most of the indoor instances. For other situations, we directly leverage the category as the prompt. We observe that Stable-Diffusion may produce some unreliable predictions which are 'too creative' and can not faithfully improve the image quality but turn it into another image, as shown in Fig. 3. To solve this issue, we generate multiple enhanced images $\{e_i^j\}_{j=1}^{M}$ for each instance image t_i and filter out the bad generation samples with the approaches described next.

https://doi.org/10.52202/079017-1234

Filter Out Bad Generation Samples. To filter out the bad generation samples produced by Stable-Diffusion and select the top K enhanced images for the following 2D-to-3D generation, we propose to leverage the CLIP models as a judge to determine which generated images $\{e_i^j\}_{j=1}^M$ from Stable-Diffusion can conform to the original appearance of the instance t_i. Specifically, we adopt the large open-sourced CLIP [47] model Open-CLIP [23] as the implementation.

We use cosine similarities $\{z_i^j\}_{j=1}^M$ between the generated instance image $\{e_i^j\}_{j=1}^M$ and the original one t_i as a metric for the selection, formulated as:

$$z_i^j = \frac{f_\theta(e_i^j) \cdot f_\theta(t_i)}{\|f_\theta(e_i^j)\| \|f_\theta(t_i)\|}, \qquad (2)$$

where f_θ is the frozen image encoder from Open-CLIP. We use the Top-K generated instance images with the largest similarities to the original t_i as the amended images. As shown in Fig. 3, Open-CLIP successfully filters out the bad generations.

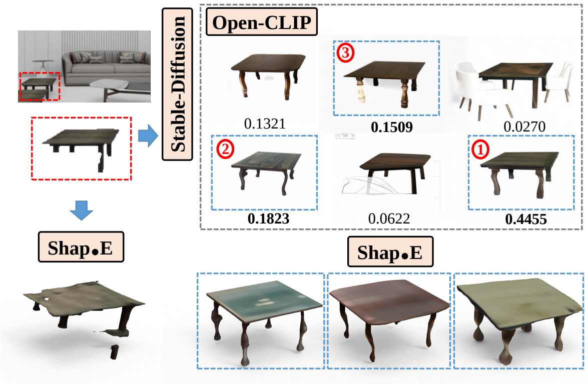

Figure 3: **Examples on the effect of our pipeline.** For the corrupted 2D instant segmented from the scene image, we leverage Stable-Diffusion to produce 6 amended generations. We then adopt Open-CLIP to filter out bad samples by judging the similarities and producing confidence scores for the generations, and keep the Top-3 generated images. The shape generations with Shap·E from the amended images are significantly more complete and accurate than the one produced by the original corrupted image.

Generate 3D models from 2D instances. The object-level 3D generation from single images [38, 61] is a well-explored task in 3D computer vision, however, most previous works show limited generation performance on open-world images. Shap·E [26] is a large model for 3D generation by training a 3D hyper diffusion model with millions of non-public 3D objects. Therefore, we leverage Shap·E to provide the deep prior to convert 2D instance images into 3D reconstructions. By this way, we obtain K 3D reconstruction proposals $\{s_i^k\}_{k=1}^K$ for each 2D instance t_i by employing Shap·E on the top K amended images.

3.3 Recovering Scene Layout

The final step is to select the most accurate 3D model proposal $\hat{s}_i$ from the K candidates $\{s_i^k\}_{k=1}^K$ and put it to the right position in a 3D scene to recover the scene layout in the input image. To achieve this, we propose a novel approach to optimize the location, orientation and size for each 3D proposal in $\{s_i^k\}_{k=1}^K$ by matching it with the estimated segmentation mask and the estimated depth. We further introduce a RANSAC-like solution for robust position optimizing. We select the reconstruction with the minimum matching error as the reconstruction $\hat{s}_i$ of t_i.

Depth Estimation. For more accurate layout estimations, we first estimate the depth map of the input image I as a 3D geometry prior. We leverage the large model Omnidata [13, 27] as the depth estimator which is trained under a collected huge depth dataset [13] containing 14-million indoor, outdoor, scene-level and object-level samples.

An issue here is that the depth D estimated for the input image I with Omnidata is not scale-aware, which can not be directly used as supervisions. We solve this problem by estimating the scale h and shift q of the predicted depth using one pair of predicted and real depth of a selected scene for each

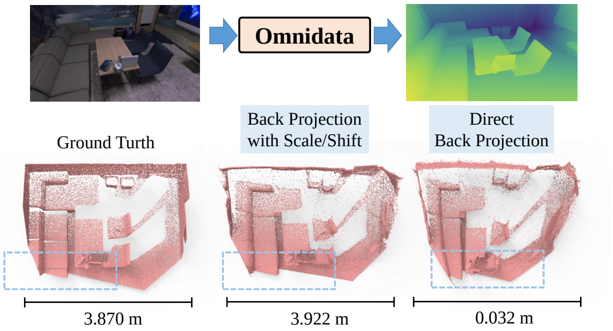

Figure 4: **Illustration of the depth transform.** The estimated depth maps from Omnidata is not scale-aware, resulting in scale inaccuracies and distortion in the back-projected depth point clouds. We achieve the accurate depth point cloud by first transforming the depth maps with the pre-solved scale and shift before back-projecting.

dataset. Specifically, we leverage least-squares criterion [14, 49] which has a closed-form solution to solve the depth scale and shift by matching the pair of predicted and real depth with a specific scene camera intrinsic parameter $\mathbb{C}_K$. After transforming D with the scale h and shift q, we can now back-project D into the 3D space with camera intrinsic parameter $\mathbb{C}_K$, achieving a 3D scene depth point cloud. The example in Fig. 4 shows that the depth point cloud produced using the transformed depth maps precisely aligns with the ground truth scene. Alternatively, we can use metric depth estimation methods [63, 66, 64, 28], which naturally handle depth scale and shift, to replace Omnidata and lessen the reliance on ground truth depth data. The depth point cloud d_i for each instance t_i is further achieved by masking the back-projected 3D point cloud.

Pose/Scale Optimization. We further estimate the scale and pose of each 3D model proposal s_i^k to put them into the right position in the 3D scene. We propose to solve this problem with an optimization-based approach on the location, rotation and size for s_i^k per the mask m_i from the Grounded-SAM and the depth point cloud d_i from Omnidata.

We model this problem as a 7-DoF shape registration task with 3-DoF of translation (tx, ty, tz), 3-DoF of rotation (rx, ry, rz) and one DoF of object scale (v). Specifically, we first sample a point cloud p_i^k from the mesh of a 3D proposal s_i^k and initialize a 7-DoF transform as a transformation function f_ϕ with learnable 7-DoF parameters ϕ. We then project p_i^k with f_ϕ to achieve the transformed prediction $\hat{p}_i^k$ by:

$$\hat{p}_i^k = f_\phi(p_i^k). \tag{3}$$

We obtain $\hat{p}_i^k$ by optimizing the 7-Dof parameters ϕ with supervisions. With the estimated depth d_i, we can draw the direct 3D matching supervision by minimizing the Chamfer Distance Loss between the transformed $\hat{p}_i^k$ and the depth points d_i. However, merely with the 3D matching constraint, the pose/scale optimization do not always converge stably since the predicted depth d_i is usually with noises in complex scenes, which significantly affects the registration on shapes.

To resolve this issue, we get inspirations from [8] to leverage the mask information predicted by Stable-Diffusion as an extra matching supervision in 2D space. Specifically, we project the transformed 3D point cloud $\hat{p}_i^k$ to the 2D space with the camera intrinsic parameters $\mathbb{C}_K$, resulting in a 2D point cloud $\tilde{p}_i^k$. Meanwhile, we form another 2D point cloud $\tilde{m}_i$ from the mask m_i by randomly sampling dense 2D points on the occupied region of the mask m_i. We then use the 2D matching constraint to minimize the 2D Chamfer Distance between $\tilde{p}_i^k$ and $\tilde{m}_i$. We illustrate the effect of 2D Matching constraint with an optimization example in Fig. 5. The final loss for pose/scale optimization of 3D reconstruction s_i^k is then formulated as:

$$\mathcal{L} = \mathcal{L}_{CD}^{3D}(d_i, \hat{p}_i^k) + \mathcal{L}_{CD}^{2D}(\tilde{p}_i^k, \tilde{m}_i). \tag{4}$$

Robust RANSAC-like Solution. With the optimization-based 7-DoF registration, we are now able to put the generated 3D object proposals into the 3D scene. However, if the mis-registration is quite large, especially in the rotation, the optimization may be trapped in a local optimum and fail to produce accurate registrations. We further introduce a RANSAC-like solution to enhance the robustness of pose/scale optimization. Specifically, we repeat the optimization r times with randomly initialized rotation matrices for f_ϕ each time. The final transform for the 3D proposal s_i^k is selected as the one with minimum matching loss in Eq. (4) among r optimal optimizations, and we define the matching error w_i^k of s_i^k as the minimal matching loss.

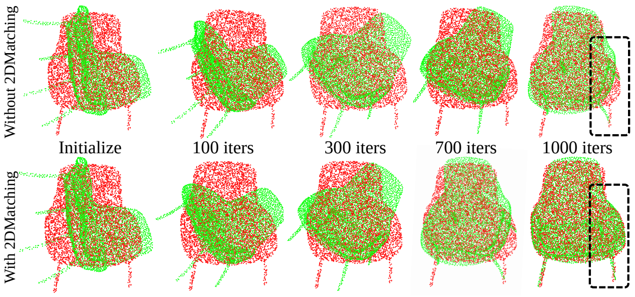

Figure 5: **Effect of the 2D Matching.** An example of optimizing the pose and scale for a chair. We visualize the optimization in 2D space. The red 2D points indicate the dense 2D point cloud sampled in the mask, which is the target. And the green 2D points donate the 2D projection of transformed 3D point clouds sampled from the generated shape of this chair instance. More robust registration is achieved with the proposed 2D matching constraint. The total 1,000 iterations take 9.2 seconds on a single 3090 GPU.

Method	3D-Front			BlendSwap			Replica		
	CDL1-S ↓	CDL1 ↓	F-Score↑	CDL1-S ↓	CDL1 ↓	F-Score↑	CDL1-S ↓	CDL1 ↓	F-Score↑
Mesh R-CNN [18]	0.449	0.471	23.90	0.265	0.406	21.87	0.268	0.408	25.42
Total3D [42]	0.198	0.520	18.44	0.133	0.400	26.93	0.390	0.780	24.01
PanoRecon [10]	0.120	**0.125**	31.94	0.355	0.417	17.11	0.326	0.440	17.13
Ours	**0.109**	0.134	**35.67**	**0.106**	**0.089**	**73.19**	**0.113**	**0.110**	**70.48**

Table 1: Comparisons on scene reconstruction from single images. Lower is better for CDL1 (i.e., Chamfer Distance), higher is better for F-Score. CDL1-S is the single-direction Chamfer Distance from the generated objects to the ground truth meshes.

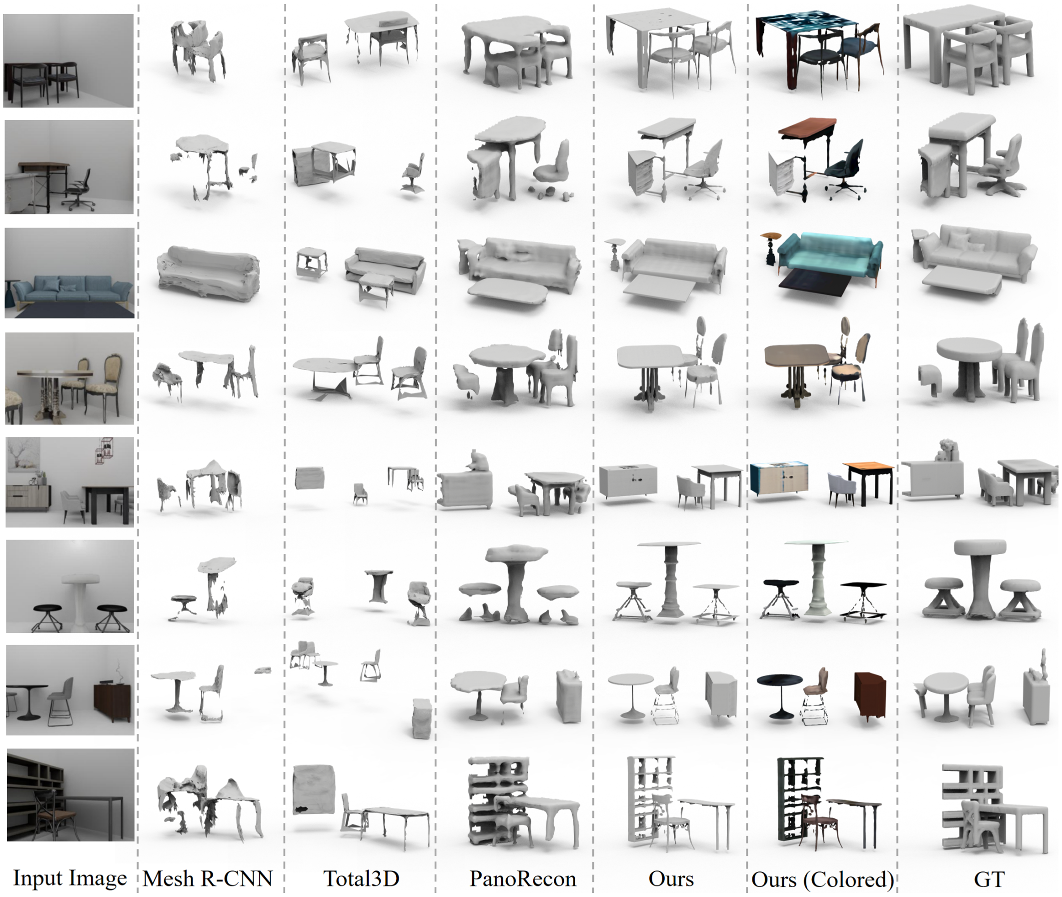

Figure 6: Comparisons on scene reconstruction from single images under the 3D-Front dataset.

We repeat the above procedure for each one of the K 3D proposals $\{s_i^k\}_{k=1}^K$. We select the 3D proposal with the minimum w_i^k as the final reconstruction $\hat{s_i}$ of t_i. The final scene generation from the single image I is achieved by combining the transformed $\{\hat{s_i}\}_{i=1}^N$ together.

4 Experiments

4.1 Setup

Implement Details. The number M of samples generated by Stable-Diffusion for each instance is set to 6, where we select the Top $K = 3$ samples with Open-CLIP. The pose/scale optimization is repeated for $r = 10$ times for each instance with RANSAC-like solution.

Datasets. We evaluate deep prior assembly under four widely-used 3D scene reconstruction benchmarks 3D-Front [17], Replica [55], BlendSwap [2] and ScanNet [11].

3D-Front [17] is a synthetic 3D dataset of indoor 3D scenes. We adopt the data pre-processed by PanoRecon [10] and randomly select 1,000 scene images from the test set as the single-image dataset. Note that all the images are captured parallel to the ground with camera locations at 0.75m height

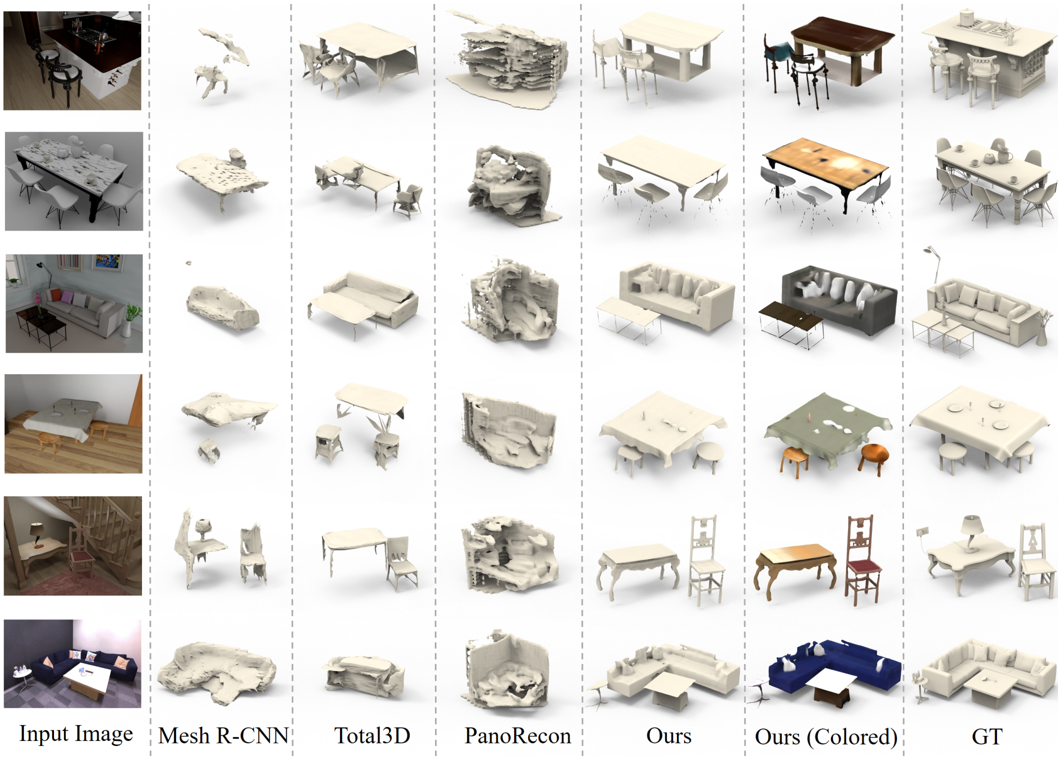

Figure 7: Comparisons on scene reconstruction from single images under Replica and BlendSwap dataset.

above the floor in the 3D-Front dataset. We follow PanoRecon to achieve the corresponding ground truth mesh for each image by only keeping the geometry at the same view and cull anything outside of the view frustum.

The Replica [55] dataset is an indoor scene dataset which contains 8 scanned 3D indoor scene with highly photo-realistic 3D indoor scene reconstruction at both room and flat scale. We adopt the pre-processed data provided by MonoSDF [67] and sample one image for each scene as the single-image dataset. The ground truth meshes are obtained with the same way as 3D-Front.

The BlendSwap [2] dataset is a high-fidelity synthesis 3D scene dataset collected by Neural-RGBD [2], containing 9 scenes with complex geometries. We collect single-view images and corresponding ground truth meshes with the same way as Replica dataset.

The ScanNet [11] dataset is a real-word 3D scene dataset captured by RGB-D cameras. We select 7 scenes from the test set of ScanNet and sample one image from each scene as the input.

Baselines. We mainly compare our method with the state-of-the-art methods in scene reconstruction from single images, i.e., Mesh R-CNN [18], Total3D [42] and PanoRecon [10]. Note that all these methods are data-driven methods and trained under 3D datasets with ground truth 3D annotations, while our method solves the task in a zero-shot manner. This means that we do not require any data-driven training on any 3D or 2D datasets, which is a much more flexible and general solution for the single image reconstruction task. We direct evaluate these methods with the official codes and the pre-trained models for numerical and visual comparisons.

Metrics. We use Chamfer Distance and F-Score with the default threshold of 0.1 following [42, 46] as metrics. Since Mesh R-CNN and Total3D only predicts the 3D objects and do not generate the backgrounds (e.g. wall and floor), we further report the single-direction Chamfer Distance from the generated objects to the ground truth meshes, i.e., CDL1-S, to only evaluate the accuracy of generated objects. Note that Total3D can generate the scene layout which can roughly represent the background, however, we find that Total3D generates layouts with large errors on all the three datasets. Therefore we do not keep the layout of Total3D for evaluation. While we achieve the background points by back-projecting the segmented background depth maps. We sample 10k points on the ground truth

https://doi.org/10.52202/079017-1234

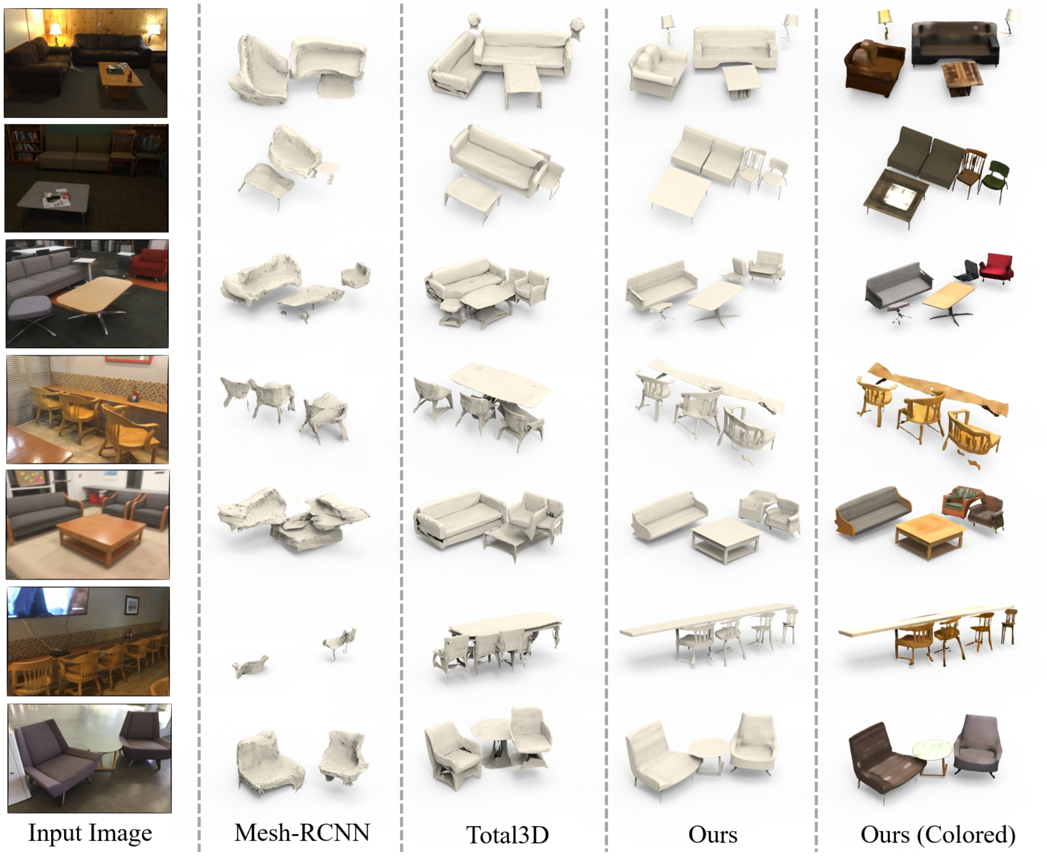

Figure 8: Comparisons of the scene reconstructions under the real-captured images from ScanNet.

meshes and the generated scenes of each methods for evaluation. Please refer to the appendix for more details on evaluation.

4.2 Scene Reconstruction on 3D-Front

Tab. 1 reports numerical comparisons on the 3D-Front dataset. We achieve the best performance among the state-of-the-art methods. Specifically, PanoRecon is trained under the 3D-Front dataset, therefore it shows convincing results in this dataset. Mesh R-CNN and Total3D are trained under Pix3D [57]/ShapeNet [7] and SUN-RGBD [54]/Pix3D datasets, respectively.

The qualitative comparison is shown in Fig. 6, where we remove the background geometries for PanoRecon, ours and GT for a clear visual comparison on the generated instances among all the methods. We further show the colored scene since the used object generator Shap·E is able to generate textured 3D objects. The visualization demonstrates our superior performance to produce accurate and visual-appealing scene reconstruction from merely a single image in a zero-shot manner.

4.3 Scene Generation of Open-World Images

We further evaluate our method under the open-world images from the BlendSwap dataset and the indoor scene dataset Replica. The quantitative comparisons in these two datasets are shown in Tab. 1, where we achieve the best performance over all the baseline methods. Note that the performance of PanoRecon [10] largely degrades under open-world scene images compared to the performance under 3D-Front dataset. The reason is that PanoRecon fails to generalize to the out-of-distribution inputs and can only handle the specific image patterns in the trained 3D-Front dataset.

The visual comparison is shown in Fig. 7, where we significantly outperform the previous works in the generation accuracy and completeness. Specifically, as shown in the 3-rd and 5-th row in Fig. 7, our method generates more accurate geometries for the table with thin legs and the chair with a

https://doi.org/10.52202/079017-1234

complex back. While Mesh R-CNN and Total3D can only generate the coarse 3D shapes and also fail to estimate accurate layout.

4.4 Scene Reconstruction from Real Images

We further evaluate deep prior assembly under the ScanNet [11] using real images. For a qualitative comparison with other methods, we select 7 scenes from the test set of ScanNet and sample one image from each scene as the input. We compare deep prior assembly with the state-of-the-art methods in scene reconstruction from single images, e.g., Mesh R-CNN [18] and Total3D [42]. We do not compare with PanoRecon [10] here since it fails to generalize to the out-of-distribution inputs and can only handle the specific image patterns in the trained 3D-Front dataset, as demonstrated in Fig. 7.

We show the visual comparisons in Fig. 8, where we successfully reconstruct scenes from real images and significantly outperform the previous works in the reconstruction accuracy and completeness. This demonstrates the huge potentials of the assembled deep priors in reconstructing real-world 3D scenes. Note that the real-world images are often blurry and corrupted when the camera doesn't focus well, e.g., the blurry input image shown in the 5-th row in Fig. 8. While our proposed deep prior assembly can also handle these challenging situations due to the powerful and robust deep priors from the large vision models.

4.5 Ablation Study

Framework Design. To evaluate the major components in our methods, we conduct ablations under the Replica dataset [55] and report the results in Tab. 2. We first justify the effectiveness of introducing Stable-Diffusion for enhancing and inpainting images as shown in 'W/o Stable-Diffusion', where we directly adopt the segmented instances for shape generation without leveraging Stable-Diffusion for enhancing and inpainting them. We then report the performance of removing the 2D or 3D matching constraints as shown in 'W/o 2D-Matching' and 'W/o 3D-Matching'. The ablation studies demonstrate the effect of each design by significantly improving the generation performance. Note that the pose/scale optimization is broken without 3D-Matching since the only 2D-Matching does not involve depth information.

Ablation	CDL1-S ↓	CDL1 ↓	F-Score↑
W/o Stable-Diffusion	0.128	0.125	67.22
W/o 2D-Matching	0.124	0.121	68.42
W/o 3D-Matching	0.199	0.168	56.08
Full	**0.113**	**0.110**	**70.48**

Table 2: Ablation studies on framework designs.

Effect of Open-CLIP and RANSAC-like solution. We further evaluate the effectiveness of filtering bad samples with Open-CLIP and the RANSAC-like solution for robust pose / scale optimization. The results is shown in Tab. 3, where both components improve the scene reconstruction accuracy.

Open-CLIP	RANSAC	CDL1-S ↓	CDL1 ↓	F-Score↑
✓	✗	0.121	0.118	69.28
✗	✓	0.129	0.123	68.92
✓	✓	**0.113**	**0.110**	**70.48**

Table 3: Ablation studies on the effect of Open-CLIP filtering and RANSAC-like solution.

5 Conclusion

We introduce deep prior assembly, a novel framework that assembles diverse deep priors from large models for scene reconstruction from single images in a zero-shot manner. This approach breaks down the task into several sub-tasks, each of which is handled by a deep prior. We do not rely on any 3D or 2D data-driven training, and provide the key solutions on layout estimation and occlusion parsing to make all deep priors work together robustly. We report analysis, numerical and visual comparisons to show remarkable performance over the latest methods.

6 Acknowledgement

This work was supported by National Key R&D Program of China (2022YFC3800600), the National Natural Science Foundation of China (62272263, 62072268), and in part by Tsinghua-Kuaishou Institute of Future Media Data.

References

[1] Motilal Agrawal and Larry S Davis. A probabilistic framework for surface reconstruction from multiple images. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021.

[4] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

[5] Adrian Broadhurst, Tom W Drummond, and Roberto Cipolla. A probabilistic framework for space carving. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 388–393. IEEE, 2001.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[8] Chao Chen, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Unsupervised learning of fine structure generation for 3D point clouds by 2D projections matching. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 12466–12477, 2021.

[9] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2023.

[10] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34:8282–8293, 2021.

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021.

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[15] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

[16] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021.

[17] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021.

[18] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9785–9795, 2019.

[19] Liang Han, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Binocular-guided 3d gaussian splatting with view consistency for sparse view synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

https://doi.org/10.52202/079017-1234

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[22] Han Huang, Yulun Wu, Junsheng Zhou, Ge Gao, Ming Gu, and Yu-Shen Liu. NeuSurf: On-surface priors for neural surface reconstruction from sparse input views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[23] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[24] Chuan Jin, Tieru Wu, Yu-Shen Liu, and Junsheng Zhou. Music-udf: Learning multi-scale dynamic grid representation for high-fidelity surface reconstruction from point clouds. *Computers & Graphics*, page 104081, 2024.

[25] Chuan Jin, Tieru Wu, and Junsheng Zhou. Multi-grid representation with field regularization for self-supervised surface reconstruction from point clouds. *Computers & Graphics*, 2023.

[26] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[27] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022.

[28] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[30] Shujuan Li, Junsheng Zhou, Baorui Ma, Yu-Shen Liu, and Zhizhong Han. NeAF: Learning neural angle fields for point normal estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[31] Shujuan Li, Junsheng Zhou, Baorui Ma, Yu-Shen Liu, and Zhizhong Han. Learning continuous implicit field with local distance indicator for arbitrary-scale point cloud upsampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[32] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[35] Baorui Ma, Haoge Deng, Junsheng Zhou, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. GeoDream: Disentangling 2D and geometric priors for high-fidelity and consistent 3D generation. *arXiv preprint arXiv:2311.17971*, 2023.

[36] Baorui Ma, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Towards better gradient consistency for neural signed distance functions via level set alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17724–17734, 2023.

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[39] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

[40] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[41] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Learning 3d scene priors with 2d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

https://doi.org/10.52202/079017-1234

Recognition, pages 792–802, 2023.

[42] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020.

[43] Takeshi Noda, Chao Chen, Xinhai Liu Weiqi Zhang and, Yu-Shen Liu, and Zhizhong Han. MultiPull: Detailing signed distance functions by pulling multi-level queries at multi-step. In *Advances in Neural Information Processing Systems*, 2024.

[44] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021.

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[46] Songyou Peng, Chiyu Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. *Advances in Neural Information Processing Systems*, 34, 2021.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[49] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[50] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[52] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001.

[53] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009.

[54] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[55] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[56] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[57] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[59] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. NeuRIS: Neural reconstruction of indoor scenes using normal priors. In *17th European Conference on Computer Vision*, pages 139–155. Springer, 2022.

[60] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021.

https://doi.org/10.52202/079017-1234

[61] Xin Wen, Junsheng Zhou, Yu-Shen Liu, Hua Su, Zhen Dong, and Zhizhong Han. 3D shape reconstruction from 2D images with disentangled attribute flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3803–3813, 2022.

[62] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

[65] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.

[66] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023.

[67] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In *Advances in Neural Information Processing Systems*.

[68] Wenyuan Zhang, Yu-Shen Liu, and Zhizhong Han. Neural signed distance function inference through splatting 3d gaussians pulled on zero-level set. In *Advances in Neural Information Processing Systems*, 2024.

[69] Wenyuan Zhang, Kanle Shi, Yu-Shen Liu, and Zhizhong Han. Learning unsigned distance functions from multi-view images with volume rendering priors. *European Conference on Computer Vision*, 2024.

[70] Wenyuan Zhang, Ruofan Xing, Yunfan Zeng, Yu-Shen Liu, Kanle Shi, and Zhizhong Han. Fast learning radiance fields by shooting much fewer rays. *IEEE Transactions on Image Processing*, 2023.

[71] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3D: A universal model for panoptic 3D scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2023.

[72] Junsheng Zhou, Baorui Ma, Shujuan Li, Yu-Shen Liu, Yi Fang, and Zhizhong Han. Cap-udf: Learning unsigned distance functions progressively from raw point clouds with consistency-aware field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[73] Junsheng Zhou, Baorui Ma, Shujuan Li, Yu-Shen Liu, and Zhizhong Han. Learning a more continuous zero level set in unsigned distance fields through level set projection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[74] Junsheng Zhou, Baorui Ma, and Yu-Shen Liu. Fast learning of signed distance functions from noisy point clouds via noise to noise mapping. *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[75] Junsheng Zhou, Baorui Ma, Liu Yu-Shen, Fang Yi, and Han Zhizhong. Learning consistency-aware unsigned distance functions progressively from raw point clouds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[76] Junsheng Zhou, Baorui Ma, Wenyuan Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[77] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring Unified 3D Representation at Scale. In *International Conference on Learning Representations (ICLR)*, 2024.

[78] Junsheng Zhou, Xin Wen, Baorui Ma, Yu-Shen Liu, Yue Gao, Yi Fang, and Zhizhong Han. 3d-oae: Occlusion auto-encoders for self-supervised learning on point clouds. *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[79] Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. Diffgs: Functional gaussian splatting diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[80] Junsheng Zhou, Weiqi Zhang, Baorui Ma, Kanle Shi, Yu-Shen Liu, and Zhizhong Han. Udiff: Generating conditional unsigned distance fields with optimal wavelet diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Appendix

A More Ablation Studies and Analysis

A.1 Alternatives on sub-tasks.

We explore the effectiveness of our chosen solutions in each sub-task by comparing them with the alternatives. Specifically, we conduct ablations to replace Shap·E [26] with One-2-3-45 [32], replace Open-CLIP [47, 23] with EVA-CLIP [56] and replace Omnidata [13] with MiDaS [50] in Tab. 4. We visually compare Shap·E with One-2-3-45 for shape generation in Fig. 9, where the results demonstrate that Shap·E is a more robust solution for generating 3D models from 2D instances.

A.2 The Effect of Instance Scale

We further conduct ablation studies to explore the effect of the instance scales to the generation qualities of Shap·E as described in **"Enhance and Inpaint 2D instances"** of Sec. 3.2 in our paper. We provide an visual comparison of the generations with different instance scales in Fig. 10. The results show that Shap·E is quite sensitive to the scale of instances in the images, where a too small or too large scale will lead to inaccurate generations with unreliable geometries and appearances. We set the scale to 6 where the Shap·E performs the best in shape generation from instance images according to our experiences.

A.3 The Effect of Confidence Threshold.

We further conduct ablations to evaluate the effect of confidence threshold σ as described in **"Detect and Segment 2D instances."** of Sec.3.2 in our paper. As shown in Fig. 11, a too large σ will drop too many instances and a too small σ struggles to filter bad instances. We choose σ = 0.4 as the suitable confidence threshold.

B More Comparisons with Data-Driven Reconstruction Methods

We additionally compare our method with SOTA data-driven scene reconstruction works PanoRecon [10], BUOL [9] and Uni-3D [71]. We show the visual comparisons under 3D-Front and ScanNet datasets in Fig. 12, where our method achieves better and more visual-appealing results under both 3D-Front and ScanNet datasets. Specifically, our method significantly outperforms other methods using real-world images in ScanNet. The reason is that all the three methods are trained under 3D-Front, and struggle to generalize on real-world images.

We further compare deep prior assembly with ScenePrior [41] in Fig.13. As shown, our method clearly outperforms ScenePrior in terms of the quality of scene geometries. Moreover, ScenePrior can only reconstruct the geometry, whereas deep prior assembly is capable of recovering high-fidelity scene appearances as well.

C Background Reconstruction

We demonstrate that deep prior assembly can also reconstruct the background geometries from the scene images. We show two scene reconstructions with backgrounds (i.e. floor, wall) in BlendSwap and Replica datasets in Fig. 14. The backgrounds are achieved by fitting planes to the projected

Ablation	CDL1-S ↓	CDL1 ↓	F-Score↑
One-2-3-45	0.123	0.122	67.98
EVA-CLIP	**0.113**	0.111	70.41
MiDaS	0.120	0.119	68.71
Ours	**0.113**	**0.110**	**70.48**

Table 4: Ablation studies on the sub-task alternatives.

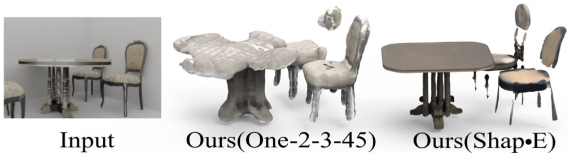

Figure 9: Ablation on shape generation alternatives.

https://doi.org/10.52202/079017-1234

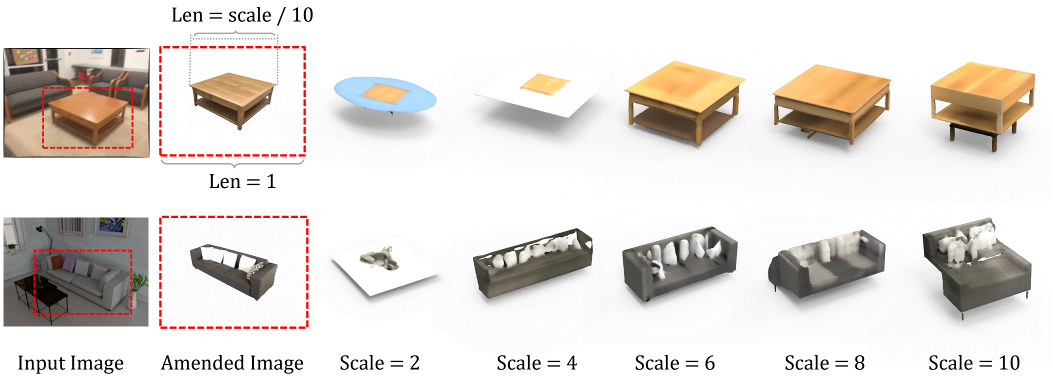

Figure 10: The ablation study on the instance scale. We select one instance for each input image and show the amended instance images. The generations obtained with Shap·E under different instance scales are visualized on the right.

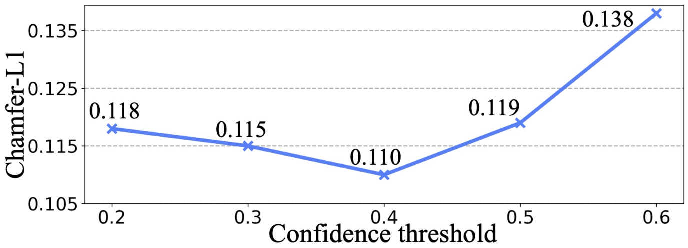

Figure 11: The ablation study on the confidence threshold.

background depth points in a similar way as our pose/scale optimization algorithm. We then cull the geometries outside of the view frustum for a clear visualization.

D Outdoor Scene Reconstruction

We further conduct experiments to evaluate deep prior assembly on complex outdoor scenes and scene containing animals, as shown in Fig. 15. The first image comes from KITTI dataset, others are collected from the Internet. With the help of powerful large foundation models, deep prior assembly demonstrates superior zero-shot scene reconstruction performance in these real-world outdoor scenes.

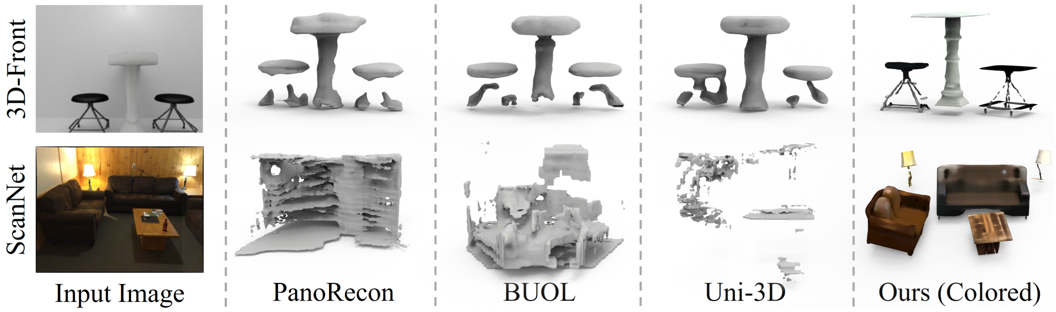

Figure 12: Visual comparisons under 3D-Front and ScanNet dataset.

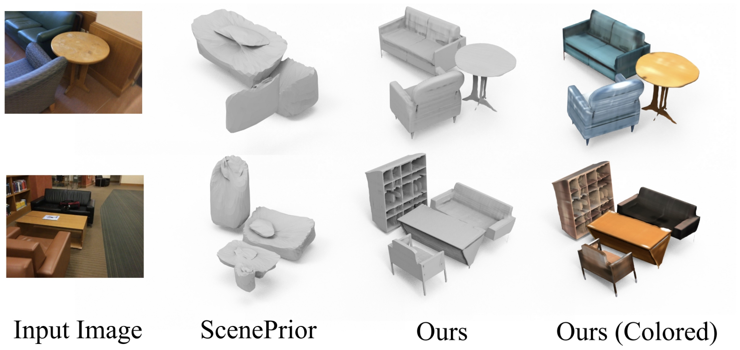

Figure 13: Visual comparisons with ScenePrior under ScanNet dataset.

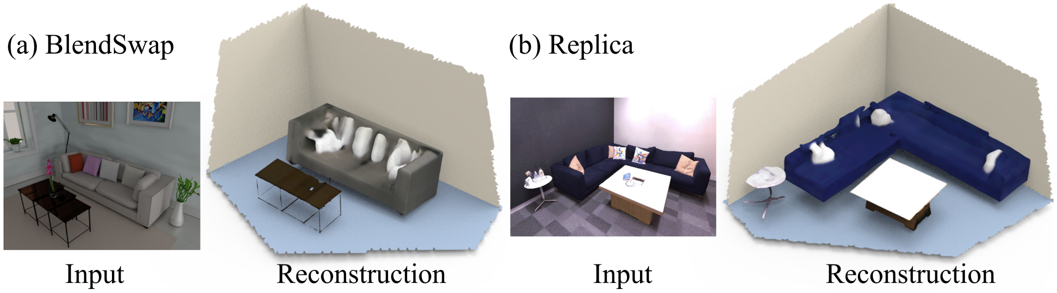

Figure 14: Scene reconstructions with backgrounds.

E Efficiency Analysis

We further evaluate the efficiency of our proposed deep prior assembly by reporting the average runtime of each sub-pipeline in our framework. The results in Tab. 5 show that reconstructing a scene from a single image takes less than 3 minutes in total, where the inference of Grounded-SAM, Open-CLIP and Omnidata takes only about 1 second. The most time consuming parts include the StableDiffusion, Shap·E and the RANSAC-like pose/scale optimization. For these three parts, we process all instances of the scene in parallel, resulting in significant time savings.

F Evaluation Details

PanoRecon. For evaluating PanoRecon [10], we adopt the official code and pretrained models for inference and directly report the performance under 3D-Front [16] dataset by evaluating the metrics (e.g. Chamfer Distance and F-Score) between the reconstructions and the ground truth meshes. For

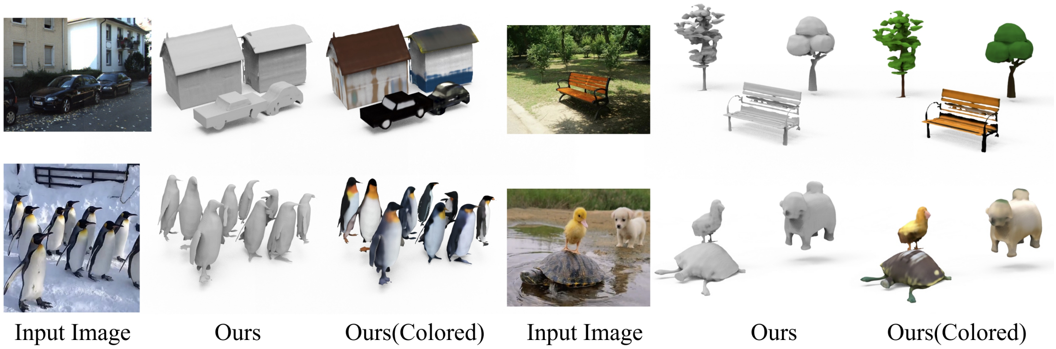

Figure 15: Outdoor scene reconstructions produced by deep prior assembly.

https://doi.org/10.52202/079017-1234

	G-SAM	Sta.-Diff.	Open-CLIP	Omnidata	Shap·E	RANSAC-Opti	Total
Time (s)	0.93	33.6	0.05	0.21	39.2	97.2	171.2

Table 5: Runtime of each sub-pipeline on a RTX3090 GPU.

the Replica [55] and BlendSwap [2] datasets where the scene location and orientation do not match the 3D-Front dataset where the PanoRecon is trained, we first normalize the center of the predicted scenes to the ground truth scenes and then register the predicted scene reconstructions to the ground truth scenes. Specifically, we first predict an initial alignment by a global registration algorithm based on feature matching [53] with RANSAC and then leverage ICP (Iterative Closest Point) registration algorithm [52] to obtain the fine registration based on the initial alignment. The metrics are reported with the registered reconstructions and the ground truth meshes.

Total3D. We leverage the official code and the pretrained models for predicting scene reconstructions with Total3D [42]. We evaluate Total3D under 3D-Front, Replica and BlendSwap with a similar way as we evaluating PanoRecon to first normalize the predicted scenes and register them to the ground truth ones before computing metrics. Total3D only predicts 3D objects and do not generate backgrounds (e.g. wall and floor). Therefore, we further report the single-direction chamfer distance from the generated objects to the ground truth meshes, i.e., Chamfer-L1 (S), to only evaluate the accuracy of the generated objects. Note that Total3D can generate the scene layout which can roughly represent the background, however, we find that Total3D generates layouts with large errors on all the three datasets. Therefore we do not keep the layout of Total3D for evaluation.

Mesh R-CNN. We adopt the official code and the pretrained models for predicting scene reconstructions with Mesh R-CNN [18]. We notice that Mesh R-CNN produces reconstructions with larger scales than the predictions of other methods and the ground truths. Therefore, we first normalize both the center and scale of the predicted scenes to the ground truth scenes and then register the predicted scene reconstructions to the ground truth scenes with a similar way as we evaluate Total3D.

Deep Prior Assembly. We evaluate our proposed deep prior assembly in the same way. We follow the same settings as we evaluate other baselines to first normalize the center of the predicted scenes to the ground truth scenes, and then register the predicted scenes to the ground truth scenes. The background points (e.g. wall and floor) of deep prior assembly are obtained by back-projecting the background depth maps, i.e., the areas where no instances exists.

G Limitation

One limitation of our method is that it may sometimes produce reconstructions with 3D instance models that are not perfectly aligned with the 2D instance segmentations in the input images. For example, the left chair generation in the last row of Fig. 8 exhibits a different color from the 2D instance of the input image. The reason is that we use Stable-Diffusion to enhance and inpaint the 2D instances, and then leverage Shap·E to generate 3D reconstructions. This process can introduce some randomness in the generated textures. The randomness primarily affects the appearances, while the geometries remain accurate. However, we justify that most reconstructions can faithfully recover the consistent scene appearances and geometries from the input images. Another limitation of deep prior assembly is that scene reconstructions may exhibit distortion due to inaccurate depth scale and shift. This issue can be addressed by replacing Omnidata with recent advances in metric depth estimation methods [63, 28].

https://doi.org/10.52202/079017-1234

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We analysis the limitations of our method in Sec.H of the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-1234

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the detailed information in reproducing our methods in Sec.3, Sec.4 of the main paper and the appendix. We also provide a demonstration code of our method in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our demonstration code as a part of our supplementary materials. We will release the source code, data and instructions upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the training and testing details in the experiment section (Sec.4) and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report the average performance as the experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-1234

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources needed to reproduce the experiments are provided in Sec.F of the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the applications and potential impacts of our method in the introduction.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-1234

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the open-sourced datasets under their licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-1234

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1234